International Journal of Mathematics and
Computer Applications Research (IJMCAR)
ISSN (P): 2249-6955; ISSN (E): 2249-8060
Vol. 7, Issue 5, Oct 2017, 1-20

TRANS STELLAR
•Journal Publications •Research Consultancy

ARIMA AND ARIMAX STOCHASTIC MODELS FOR FERTILITY IN NIGERIA

ADEKANMBI DAMILOLA BOLANLE[1] & AKINYEMI OLUWADARE[2]

[1]*Ladoke Akintola University of Technology, Ogbomoso, Nigeria*

[2]*Ekiti State University, Ado-Ekiti, Nigeria*

ABSTRACT

The aim of this study to compare forecasting abilities of two time series models: Univariate autoregressive integrated moving average (ARIMA) and autoregressive integrated moving average with exogenous variable, (ARIMAX). A stochastic time series model for live births series in Nigeria was built, starting from an identified univariate ARIMA model. The first step in formulating the ARIMAX model for the series was to identify a suitable ARIMA model for such series. ARIMAX time series model is a healthy marriage between regression and ARIMA model. A univariate ARIMA (1, 1, 0) model was developed for the disaggregated live births series, and was found adequate in modelling the series, as confirmed by the results of the diagnostic checks conducted on the model. One of the demographic factors that could have influence on livebirths is the population of women-of-childbearing-age, which was the exogenous variable in the ARIMAX model. The cross-correlation function (ccf) of the bivariate process of the livebirths series and women-of-child-bearing-age series gave an indication that, live births were related to the current and previous lagged values of the predictor variable. Inclusion of the exogenous variable into the identified ARIMA model yielded ARIMAX (1,1,0,1) model. The results of the measures of model adequacy and forecast accuracy suggested that both the ARIMA and the ARIMAX models have satisfactory predictive ability for the live births series. The ARIMAX model was considered to be a more suitable model due to its slightly smaller AIC with better MAPE compared with the ARIMA model. It may be true that live births are also influenced by factors other than women-of-childbearing-age, but the inclusion of this exogenous variable in the identified ARIMA model captured the major variations in live births in Nigeria. Forecast of future live births in Nigeria will aid in determining forecast demands of this demographic phenomenon on the various systems in the country.

KEYWORDS: Autoregressive Integrated Moving Average (ARIMA) Model, Autoregressive Integrated Moving Average (ARIMAX) Model With Exogenous Variable, Autocorrelation Function (Acf), Partial Autocorrelation Function (Pacf) & Cross- Correlation Function (Ccf)

Received: Aug 01, 2017; **Accepted**: Aug 21, 2017; **Published**: Oct 03, 2017; **Paper Id.**: IJMCAROCT20171

Original Article

INTRODUCTION

Univariate time series models have been found suitable for evaluating short-term effects of demographic variables, [29, 30, 31, 36, 39]. A time series is a stochastic process, which is a collection of random variables ordered in time, where the time index *t* takes on a finite or countable infinite set of value, [9, 13, 18]. Time series models such as autoregressive moving average (ARMA), autoregressive integrated moving average (ARIMA), and seasonal autoregressive integrated moving average (SARIMA) are referred to as univariate time series models because, only one variable that depends on its past values is included in the models. There are issues of concern as regards the adequacy of the univariate models in modelling demographic variables. Empirical specification of univariate models could be hampered, by fluctuations that can be attributed to one or more variables other than

one variable usually considered in such models. McDonald [30] pointed out that the models are usually characterized by poor performance at their turning points and consequently resulting in poor forecasts. Part of the limitations of univariate stochastic time series models are the inability to predict how dependent variable changes according to some relevant explanatory variables. In order to overcome this limitation, ARIMAX model could be used which accommodates explanatory variables to explain the variations the response variable.

Univariate time series models use only past values of the variable under consideration, to forecast future values. Related explanatory variables can be inserted into the such univariate model, which results in ARIMAX time series models called autoregressive integrated moving average with exogenous variable to obtain a better forecast. Apart from improving the accuracy of forecasts, joint modelling of such series could facilitate a better understanding of the dynamic relationship among the variables, [40]. ARIMAX not only allows the effects of covariates or predictors on the response variable to be measured, but also the lagged effects of such converts. ARIMAX time series models and their applications have been discussed by many authors, [1, 8, 9, 11, 19, 22, 23]. Univariate ARIMA and ARIMAX models are proposed in this study, to capture the trend of livebirths in Nigeria and also to determine the effect of the population of women-of-childbearing-age on birth trend.

Theoretically, most demographic variables depend on a number of stochastic elements. The focus of this study is to determine whether incorporating factor that have influence on a demographic variable into a univariate time series model of such variable could lead to improved forecasts of such variable. The Population of *women-of-childbearing-age* is one of the possible predictors could have a significant relationship with livebirths, and could be useful for forecasting future livebirths of a population. For ARIMAX models, the dependent variable is explained by past values of the random variable, random shocks and explanatory variables. In this study, a univariate ARIMA model for the livebirths series will be identified, as well as an ARIMAX model of the livebirths series to model the association between live births and the number of *women-of-childbearing-age*.

In demographic context, fertility is the actual birth performance, while nasality represents the role of births in population change and human reproduction, [38]. The terms *natality*, *fertility* and *births* may relate to total births, including live births and stillbirths, but they have been used increasingly to refer to live births only, [26, 28, 38, 30]. It has been established that the crude birth rate for Nigeria stood at 36.1 in 2010, [44]. The estimated total fertility rate per woman in Nigeria has also declined from 7.2 in 1960 to 4.0 in 2013, [44]. This is an indication that though the number of live births is on the increase in Nigeria, nevertheless there is a sharp reduction in the number of children a woman between ages 15-49 would have during her reproductive life, if for all of her childbearing years she were to experience the age-specific birth rates for that given year. Between 1980 and 2003, it was observed that the birth-rate among Nigerian women aged 15-19 has declined by 27%. However, because Nigeria population increased rapidly, the annual number of births to teenage women increased by 50% over this period, [35]. Adolescent birth-rate therefore contributes significantly to the total birth rate of women in Nigeria. Nigeria has a population close to 124 million in 2003; and is one of the 10 most populous countries in the world, and the most populous country in Africa. The decline in the fertility rate of Nigerian women could be due to the increase in contraceptive usage among women-of-child-bearing-age, [44].

The National Population Commission in Nigeria is an agency with the statutory mandate to establish and maintain a uniform system of vital registration for the nation with a view of providing vital statistics on a regular basis for the purpose of socioeconomic planning, [34]. The National Population Commission reported a total of 9, 936,221 registered

live births in Nigeria during the period 1994-2007. In the recent, age-at-marriage of females has increased which consequently lead to a decrease in fertility rate in Nigeria. A decrease in total fertility rate may be due to postponement of births among the females of reproductive age in the country. The estimated crude birth rate of Nigeria, which stood at 46.2 in 1970, has reduced to 41.5 as at 2012, [42]. Data on women-of-childbearing-age were mainly extracted from the national-level reproductive health survey in Nigeria carried out by the National Population Commission in collaboration with ORC Macro. [34]. The data were recorded on an annual basis from 1994 to 2007.

For the data to be suitable for time series modelling it must cover a relatively long period and must have a high caseload. The higher the caseload, lower the possibility of the data to be influenced by random fluctuations, [9, 10, 13, 36, 18]. The data were disaggregated so as to have data with high caseload that will not be affected by random fluctuations; and to uncover the possible pattern in the live births and women-of-childbearing-age series. Boot-Feibes-Lisman first difference, a non-model based method (BFL-FD) developed by Boot *et al*, [5] which is suitable in desegregating annual time series data into quarterly figures, was employed to desegregate the annual data on record live births and *women-of-childbearing-age*

UNIVARIATE TIME SERIES MODELS

Time series modelling can contribute to understanding the physical system by revealing something about the physical process that builds persistence into a series. Univariate time series models use only past history of individual series being modelled, [9, 10, 12]. They do not use any information from other series that may be related to the series being modelled. Such prediction can be used as a baseline to evaluate the possible importance of other variables to the process.

ARMA Models

Autoreggressive moving average (ARMA) model can be used, to predict the behavior of a stationary time series of past values alone. The autoregressive terms of a variable are the lagged values of the variable that have a statistically significant relationship with its most recent value, while the moving average terms are the residuals or lagged errors resulting from previously made estimates, [9, 10, 12, 13, 14]. The attractiveness of the ARMA model is that it is a parsimonious representation of a stationary stochastic process. The major advantage of this model is its flexibility in incorporating explanatory variable which might be useful in predicting the response variable. The general form of the ARMA (p, q) scheme in lag operator notation is, [9, 10, 12, 13]:

$$\varphi(B)y_t = \delta + \theta(B)\varepsilon_t \qquad t = 1,2,.....,T \tag{1}$$

Equation (1) can be written explicitly as

$$y_t = \delta + \varphi_1 y_{t-1} + ...+ \varphi_p y_{t-p} + \varepsilon_t - \theta_1\varepsilon_{t-1} - .....\theta_q\varepsilon_{t-q} \tag{2}$$

where

B **:** Backward shift operator such that $B^j y_t = y_{t-j}$

$\varphi(B)$ **:** The autoregressive (AR) operator, such that:

$$\varphi(B) = 1 - \varphi_1 B - .... - \varphi_p B^p \ .$$

$\theta(B)$**:** The moving average (MA) operator, such that:

$$\theta(B) = 1 - \theta_1 B - .... - \theta_q B^q$$

ε_t**:** A white noise sequence also referred to as disturbance term such that

$$E(\varepsilon_t) = 0;\ E(\varepsilon_t^2) = \sigma_\varepsilon^2 \text{ and } E(\varepsilon_t \varepsilon_{t+k}) = 0.$$

δ**:** **C**onstant term.

p: Order of non-seasonal autoregressive (AR) term.

q: Order of non-seasonal moving average (MA) term.

Equation (1) is referred to as an *autoregressive moving average* (ARMA) *process*. Simply written as ARMA(p, q) model. The orders (p,q) of an ARMA (p,q) process are the highest lags of y_t and ε_t, respectively.

ARIMA Models

Autoregressive Integrated Moving Average (ARIMA) model is an extension of ARMA class in order to include more realistic dynamics, in particular, non-stationarity in mean and variance, [9]. ARIMA model can be considered as a special case of regression model in which the dependent variable has been stationarised and the explanatory variables are all lagging of the dependent variable and or lags of the errors. Staionarity tests, such as Augmented Dickey-Fuller test can be used to determine, whether a series is stationary or not, and consequently determining if differencing term should be included in the model specification. ARIMA process generates nonstationary series, that is integrated of order *d*, denoted *I (d)* so that an integrated series should be differentiated until it is stationary before modelling. A nonstationary *I (d)* process is one that can be made stationary by taking *d* differences. Such processes are referred to as *difference-stationary* or *unit-root* process. The difference series can then be modelled as a stationary ARMA (p, q) process. In practice many demographic series are nonstationary in mean and they can be modelled only by removing the nonstationary source of variation, which is often done by differencing the series. ARIMA models which are capable of describing a wide class of non-stationary time series containing stochastic trends has become popular univariate time series models for forecasting demographic variables, [9, 30, 31]. The models could be adopted for short term forecasts of live births of a population to determine the future live births of such population. The general form of ARIMA model is [3, 9, 10, 12, 13]:

$$\varphi(B)\nabla^d y_t = \delta + \theta(B)\varepsilon_t \tag{3}$$

Where

∇: Differencing operator, such that $\nabla y_t = y_t - y_{t-1}$ and $\nabla^d y_t = \nabla(\nabla^{d-1} y_t) = w_t$

$$\nabla = \nabla_1 = 1 - B$$

$\varphi(B)$ and $\theta(B)$ are the lag polynomials for AR and MA respectively.

(3) is referred to as ARIMA scheme of orders *p, d*, and *q* or ARIMA(p,d,q). When the order of differencing is greater than zero, i.e *d*>0, then the drift element (δ)corresponding to a d^{th}-order polynomial could be set to zero, i.e $\delta = 0$

If $d = 1$, (3) now becomes

$$\nabla y_t = \varphi \nabla y_{t-1} + ... + \varphi_p \nabla y_{t-p} + \varepsilon_t - \theta \varepsilon_{t-1} - ..... - \theta \varepsilon_{t-q} \quad (4)$$

There are three main steps in finding an appropriate model for a series, which are modelled specification or identification, estimation of model parameters and model verification. Visual inspection of the time plot, autocorrelation function (*acf)* and partial autocorrelation function (*pacf)* of a series may be employed to identify an appropriate model for such series. The specification should be guided by the principle of parsimony, by which the best model is the simplest possible model that adequately describe the data. The stationary condition of a series must be verified before identifying an appropriate model for such series. The coefficients of the identified ARIMA model can be estimated by the method of maximum likelihood or least squares regression. Finally, the last stage is model checking which involves verification of the quality of the identified model, by analyzing the model residuals to verify the randomness of the residuals. In the absence of no inadequacies, the model is considered suitable for the data. Statistical theory and applications of ARIMA models are fully discussed in [9, 10, 12, 13, 16, 18, 20, 23, 25, 26].

TIME SERIES MODEL WITH EXOGENOUS VARIABLE (ARIMAX)

ARIMAX time series models are extensions of the ideas of univariate time series, stochastic models, and are basically a healthy marriage between regression and ARIMA, [22, 25]. Empirical specification of univariate time series model may be hampered by fluctuations, that can be attributed to variables other than y_t. For ARIMAX time series model, the dependent variable is explained not only by past values of forecast variable, but also on random shocks and exogenous variables. The models are different from, transfer function models where the effects are uni-directional, and also allow for the lagged effects of covariates and for decaying effects of the covariates, [1, 18, 22, 23].

ARIMAX Model

Autoregressive Integrated Moving Average (ARIMAX) with exogenous variable is a generalization of an ARIMA model for a single series. ARIMAX model is an ARIMA model that incorporates information provided by leading indicators and other covariates. It is a combination of autoregressive model (using previous states), moving average model (using past residuals); and ordinary regression model (using external variables on integrated series). It is a special case of transfer function model, popularized by Box and Jenkins, [9]. ARIMAX model is an extension of the ARIMA model, where external covariates or exogenous variable may be added depending on cross-correlations between the response variable and the covariates. There are different names for ARIMA models with regressor series, such as ARIMAX, ARIMAX model or XARIMAX model, or an ARIMA model with regressors, [23]. ARIMAX model uses information from other related variables, to attempt to obtain reasonable forecasts of the dependent variable. The model could be used to check if a set of explanatory variable has an effect on a linear time series. Explanatory variables can be inserted into a univariate model, so that the dependent variable y_t depends on lagged values of the independent variable, and derive ARIMAX model. The dependent variable is therefore explained by past values, random shocks and explanatory variables. The identified ARIMAX model used in this study was developed based on the method proposed by Tiao and Box [40],

which is similar to the Box and Jenkins method, [9] for univariate models, except that cross-correlations between the series model for. For one dependent variable and one explanatory variable, the mathematical representation of ARIMAX *(p, d, q, b)* model has the form, [40]:

$$y_t = \beta x_t + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + ... + \varphi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - .... - \theta_q \varepsilon_{t-q} \quad (5)$$

Where

ε_t is a white noise process.

x_t is a covariate at time t.

β is the coefficient of the covariate.

p: is the order of non-seasonal AR terms

q: is the order of non-seasonal MA terms.

b: is the number of explanatory variable included in the model.

y_t: livebirths at time t.

The explanatory variable can be inserted into the univariate model, to derive the ARIMAX model. An ARIMAX model simply adds in the covariates, on the right hand side of the model. One disadvantage of ARIMAX model is that the covariate coefficient is difficult to interpret. The presence of lagged values of the response variable on the right hand side of (5) is an indication that β can only be interpreted conditional on the value of previous values of the response variable, which is hardly intuitive The general form of ARIMAX *(p,d,q,b)* model for one explanatory variable has the following condensed form in lag operator notation

$$y_t = \beta x_t + \varphi(B)^{-1} \theta(B) \varepsilon_t \quad (6)$$

The model can also be written as:

$$y_t = \frac{\beta}{\varphi(B)} x_t + \frac{\theta(B)}{\varphi(B)} \varepsilon_t \quad (7)$$

where

$$\varphi(B) = 1 - \varphi_1 B - .... - \varphi_p B^p$$

$$\theta(B) = 1 - \theta_1 B - ..... - \theta_q B^q$$

For more than one explanatory variable, the mathematical form of ARIMAX model has the form:

$$y_t = \beta x_t + \beta_1 x_{t-1} + ... + \beta_j x_{t-j} + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + ... + \varphi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - .... - \theta_q \varepsilon_{t-q} \quad (8)$$

Thus, ARIMA models allow polynomial regression when appropriate, so that an ARIMAX model is an ARIMA

model with regressors. In backshift operators, the general form of ARIMAX models with more than one explanatory variable is:

$$y_t = \sum \beta_j x_{t-j} + \varphi(B)^{-1}\theta(B)\varepsilon_t \quad (9)$$

Equivalently, the model can be written as:

$$y_t = \sum \beta_j x_{t-j} + \frac{\theta(B)}{\varphi(B)}\varepsilon_t \quad (10)$$

ARIMAX model has the same stationarity requirements as the univariate models. The response series is stationary if the roots of the homogenous characteristic equation of the form:

$$\varphi(B) = 1 - \varphi_1 B - \varphi_2 B^2 - ..... - \varphi_p B^p = 0 \quad (11)$$

Lie Outside the Unit Circle

If the response variable y_t is not stationary in mean, it should be differenced to form a stationary ARIMA model, before including explanatory variables. The series of the response variable could also be subjected to Box-Cox transformation to induce stationarity in variance, if the original series is not stationary in variance. If the explanatory variable (x_t) is not stationary, then a false negative rate for significance tests of β can increase. The staionarity transformations in mean and variance, also apply to the explanatory variable to ensure stationarity in mean and variance, if the series of the variable is not stationary. The order of differencing to apply to the explanatory series, as well as the number of explanatory variable lags to include in the model can be determined from the cross-correlation function. Differencing the explanatory series also changes the practical interpretation of β . Subsequently, β could be interpreted as expected effect a unit increase in the explanatory variable has on the difference between current and lagged values of the response variable y_t, conditional on those lagged values. In the context of this study, ARIMAX model will be employed in measuring how population of *women-of-childbearing-age* x_t affects the number of *livebirths* in Nigeria.

Cross-Correlation Functions

In the relationship between two time series y_t and x_t, the series y_t may be related to past lags of x series. The cross-correlation function (*ccf*) of two series is the product moment correlation, as a function of lag between the series; and is helpful in identifying lags of the x-variable that might be useful predictors of y_t, [12, 13]. The cross-correlation between two time series describes the normalized cross-covariance function. Cross-correlation function (*ccf*) is therefore, a basic exploratory tool, that could be employed in the identification process of ARIMAX model, and generally for time series with exogenous variable. In order to determine whether it will be necessary to add an explanation and even to determine the number of explanatory lags to include in ARIMA model which will result into ARIMAX model, the cross-correlations must be checked.

The *ccf* is a generalisation of the *acf* to the multivariate case so that, cross-correlogram analysis is an extension of correlogram analysis. The structure of *ccf* could be used, to find linear dynamic relationship in time series data that have been generated from stationary process, [14]. For non-stationary series, the estimated autocorrelation and cross-correlation functions of the (X_t, Y_t) series will fail to damp out quickly, [9]. Stationarity can therefore, be induced in such series by

differencing. It is assumed that the required degree of differencing *d* to induce stationarity has been achieved when the estimated autocorrelation and cross-correlation functions die out quickly. Threoretically, the h^{th} cross-correlations between two covariance stationary series $\{y_t\}$ and $\{x_t\}$ are defined as:

$$\rho_{y\,x,h} = \frac{E(y_t x_{t+h}) - \mu_y \mu_x}{\sqrt{V(x_t)V(y_t)}} = \frac{\gamma_{yx}(h)}{\sigma_y \sigma_x} \tag{12}$$

and

$$\rho_{xy,h} = \frac{E(x_t y_{t+h}) - \mu_x \mu_y}{\sqrt{V(x_t)V(y_t)}} = \frac{\gamma_{xy}(h)}{\sigma_y \sigma_x} \tag{13}$$

where

(X_t, Y_t) represent a pair of stochastic processes that are jointly wide sense stationary.

μ_x and σ_x are the mean and variance of the process X_t and are constant over time due to stationarity.

μ_y and σ_y are the mean and variance of the process Y_t and are also constant over time due to stationarity.

$\rho_{y\,x,h}$ or $\rho_{xy,h}$ will vary between -1 and 1. A value of $\rho_{xy,h}$ =1 indicates that, at the alignment *h*, the two time series have the exact same shape, though the amplitude may be different. A value $\rho_{xy,h}$ =-1 indicates that, they have the same shape except that, they have opposite signs. A $\rho_{xy,h}$ =0 shows that, they are completely uncorrelated. In practice, a correlation coefficient greater than 0.7 or 0.8, indicates a strong correlation. The order of the subscript is important in the notation, so that $\rho_{y\,x,}(h) \neq \rho_{xy,}(-h)$. Unlike autocorrelations, cross-correlations are not symmetric, so that the order xy or yx matters. The order of the indices indicates the variable that is measured using contemporaneous values and the variable that is lagged.

Time series often exhibit nonstationarity behaviour. A very slow decay of the *ccf* indicates non-stationarity of the series. A peak in a *ccf,* followed by a tapering pattern is an indicator that, lag 1 and possibly lag 2 values of the y-variable may be possible indicators. The *ccf* can therefore, be used to determine the number of explanatory variable lags to be included in the model. The *ccf* is helpful in identifying lags of the explanatory variable, that might be useful predictors of response variable. It is therefore, important to determine whether the series under consideration are stationary or not before using them in an ARIMAX model. Linear combinations of the elements of y_t may be stationary, and differencing all series simultaneously can lead to complications in model fitting, [7, 24]. The method to test the stationarity of the series is the unit-root test referred to as Augmented Dickey-Fuller test.

In *R* statistical computing software, the *ccf* is defined as the set of correlations between x_{t+h} and y_t, for $h = 0, \pm 1, \pm 2, \pm 3$ and so on. A negative value for h is a correlation between the x-variable at a time before *t* and the y-variable at time t, so that when $h = -2$, the *ccf* value will give the correlation between x_{t-2} and y_t. When one or more

x_{t+h} with h positive, are predictors, it said that x lags y. After the identification procedures have given rise to the selection of a particular ARIMAX model, model parameters can be estimated by method of least squares or maximum likelihood.

DIAGNOSTIC TESTS FOR TIME SERIES MODELS

After building either univariate ARIMA(p,d,q) or ARIMAX(p, d, q, b) model, it is necessary to test whether the model is adequate for the series. Different diagnostic tests can be performed, to ascertain the adequacy of the model. One of such tests is to verify if the residuals of the proposed time series model is a white noise process. The significance of the autocorrelation of the residuals could be checked if they are within the non-significance bound, which is the two standard error bounds $\pm 2/\sqrt{N}$, [26]. The adequacy of the model should be questioned if the autocorrelations of the residuals of the first N/4 lags are close to the critical bounds.

Ljung-Box statistic also referred to as Q-statistic, for high-order serial correlation is another diagnostic test for time series model. Ljung-Box statistic could be used to check randomness in residuals. If the residual is random, the model is considered reasonable for the series. The formula for the Q-statistic is:

$$Q = n(n+2)\sum_{k=1}^{m}(n-h)^{-1}\left[\rho_{\tilde{a}}(h)\right]^2 \tag{14}$$

where

$\rho_{\tilde{a}}$ **:** Are the autocorrelations of estimation residuals.

h: A prefixed number of lags.

n: Sample size.

ACCURACY OF FORECAST MODELS

Empirical evaluations of the performance of forecasting models rely upon measures of error criterion such as Mean Square Error (MSE), Mean Absolute Error (MAE) or Mean Absolute Percentage Error (MAPE). The measures are used in measuring the accuracy of the projection of the models. They provide indication on how well a model fits a series, [26]. Given a particular model for a process,

$$y_t = \mu_t + \varepsilon_t \tag{15}$$

The error after y_t is observed is $e_t = y_t - \mu_t$.

where

μ_t **:** Is a known function of past y_t and ε_t values.

e_t **:** Is a realisation of random variable ε_t and are independent and identically distributed with mean 0 and variance σ^2.

$$MSE = E\left[(y_t - \mu)^2\right] = E\left(e_t^2\right) \tag{16}$$

The Mean Absolute Error (MAE) could be computed using the formula

$$MAE = E\left[|y_t - \mu_t|\right] = E\left(|\varepsilon_t|\right) \tag{17}$$

The Mean Absolute Percentage Error (MAPE):

$$MAPE = E\left\{\frac{|y_t - \mu_t|}{y_t}\right\} \tag{18}$$

Given that y_t takes only positive values.

MODEL SELECTION

Once the appropriate time series models have been fitted, Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) can be used to choose the best between the proposed models. Akaike, [2] proposed an information criterion of the form

$$AIC = -2In(L) + 2k \tag{19}$$

where

K: Number of parameters in the model to be estimated.

L: Likelihood function of the model.

When there are two or more competing models, the model with the smallest AIC is deemed best in the sense of minimizing the forecast mean square error, [26]. It was however, pointed out by Schwartz [37] that, AIC is not a consistent criterion due to the fact that, it does not select the true model with probability approaching 1 as $n \to \infty$, [15, 25]. Schwartz [37] therefore, proposed the Bayesian Information criterion (BIC)

$$BIC = -2In(L) + kIn(n) \tag{20}$$

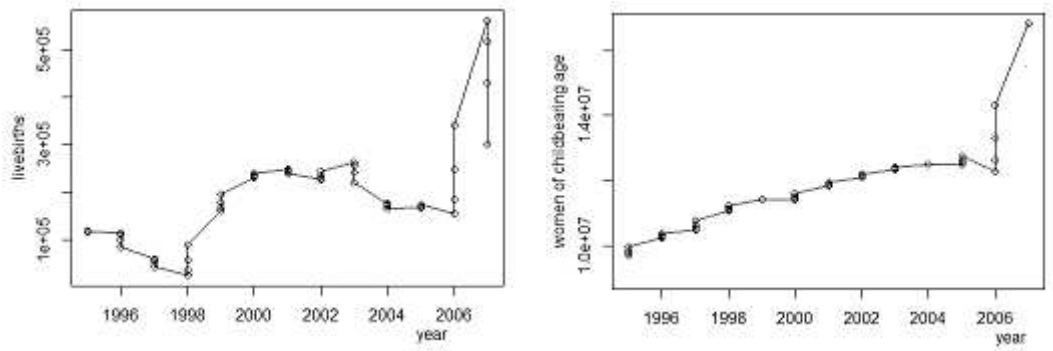
where

n: number of observed data points or sample size.

RESULTS

The time plots of the disaggregated *livebirths* series (y_t) and disaggregated *women-of-childbearing-age* series (x_t) are shown in figures 1a and 1b, respectively. In both cases, the series exhibit nonstationarity in mean, as well as non-stationarity in variance. The live births series exhibits a sudden peak around 2007 and a clear upward trend, while the *women-of-childbearing-age* series shows a noticeable sudden peak also around 2007. The correlograms of the *acf* and *pacf* of the livebirths series are depicted in figure 2a and 2b. It is obvious that, the *acf* shows an exponential decay, an indication that the two series are not stationary in mean. The Augumented Dickey-Fuller test (ADF) tests applied to the original live births series and the women-of-childbearing-age series differently also confirm the non-stationarity status of the two series. The two series are therefore subjected to Box-Cox variance stabilising transformation to remove non-stationarity in

variance in the series, and are later subjected to non-seasonal differencing, to remove non-stationarity in mean; after which they become stationary as confirmed by the ADF test.

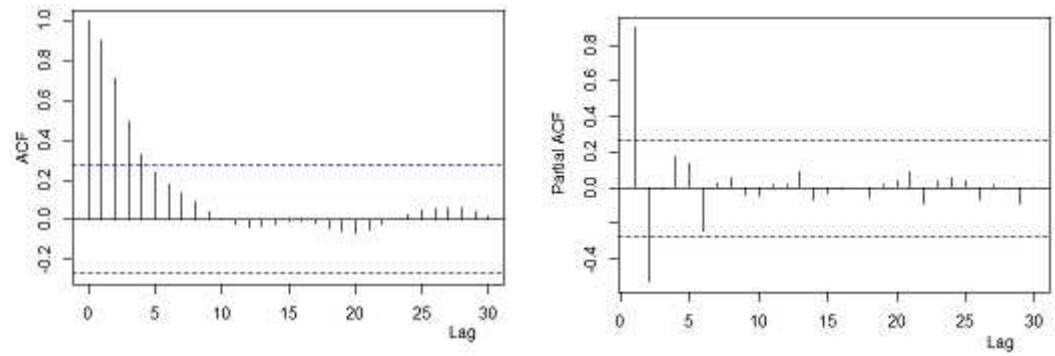


(1a) **(1b)**

Figure 1: Time Plots of the Live births Series (Y_t), and the Women-of-Childbearing-Age Series (X_t)

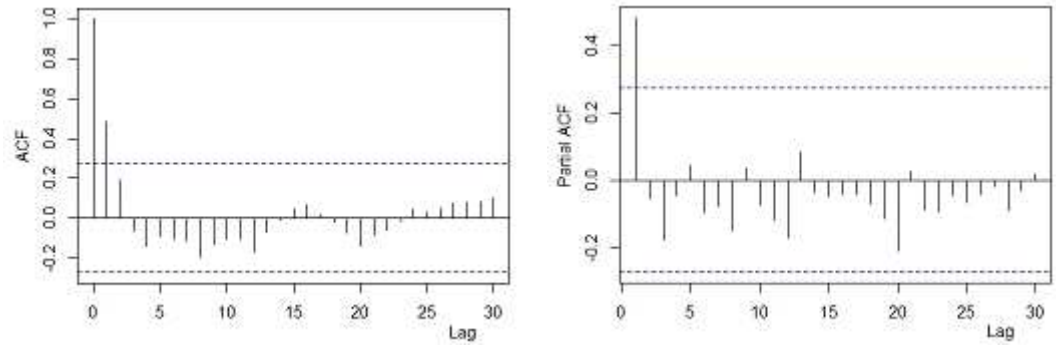
Results of the Analysis of the Univariate ARIMA Model

The *acf* of the livebirths data shown in Figure 2(a), has a sinusoidal pattern that tails off to zero reflecting the meandering shape of the series, which indicates that the series is non-stationary; suggesting that at least one non-seasonal differencing will be appropriate. The livebirths series is expected to possess non-stationarity characteristics, since the series represent 'uncontrolled' behaviour of certain demographic process outputs. Spikes in the *pacf* could indicate possible non-seasonal AR terms. The *pacf* of the series shows a clear single positive spike at lag 1 and a negative spike at lag 2. After subjecting the series to variance stabilizing transformation so that $z_t = \log(y_t)$ and non-seasonal differencing (∇z_t) to remove trend nonstationarity; the series becomes a stationary process, as shown in figures 3(a) and 3(b). The *acf* cuts off after lag 2, and the rest of the values randomly oscillate about zero, within the non-significance limits. The tapering pattern in the lags of the *acf* suggests that, a non-seasonal AR (1) may be a useful part of the model. The non-significance spikes show some patterns of groups of positive and negative values. The autocorrelations at lag 1 and 2 are positive, an indication that the AR is positive. The identified univariate model is therefore, ARIMA (1,1,0). The *acf* and *pacf* of the predictor variable are shown in figures 4(a) and 4(b), respectively.



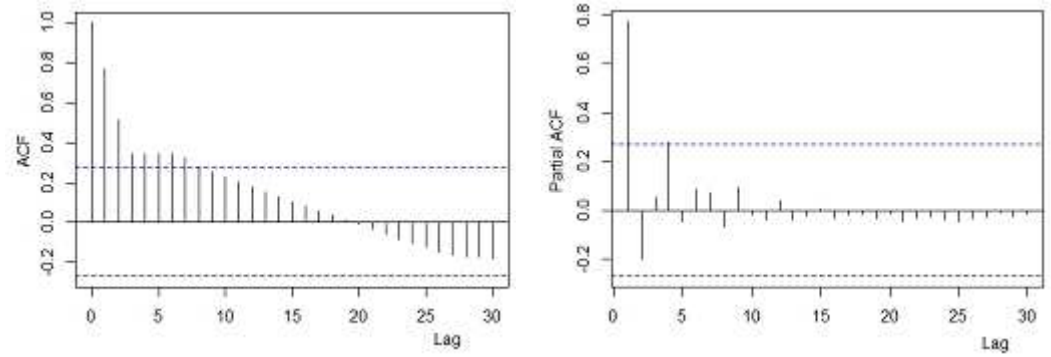
2(a) **2(b)**

Figure 2: *Acf* and *Pacf* of Livebirths Series (Y_t)



3(a) **3(b)**

Figure 3: *Acf* and *Pacf* of (∇z_t)



4(a) **4(b)**

Figure 4: Acf and Pacf of Women-of-Childbearing-Age Series (Xt)

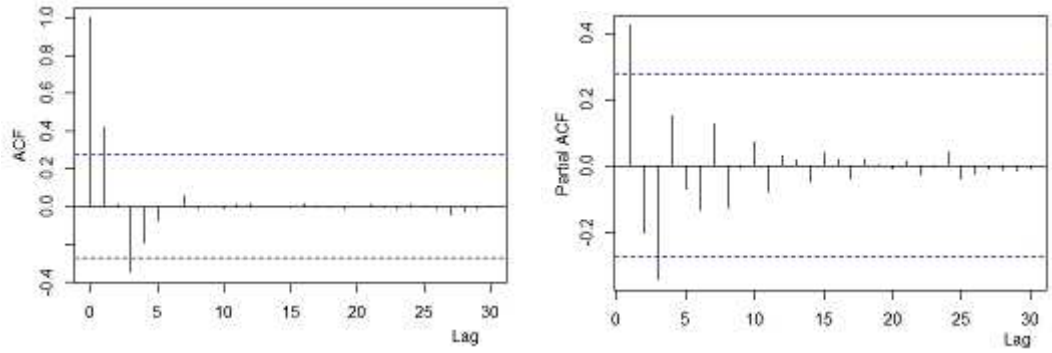


Figure 5: Acf and Pacf of ($Logx_t$) with Nonseasonal Differencing

The equation of the identified ARIMA (1,1,0) model is therefore:

$$(1-\varphi B)\nabla z_t = \varepsilon_t \tag{21}$$

The ARIMA (1,1,0) can be written explicitly as:

$$z_t - z_{t-1} - \varphi z_{t-1} + \varphi z_{t-2} = \varepsilon_t \tag{22}$$

$$z_t - z_{t-1} - \varphi(z_{t-1} - z_{t-2}) = \varepsilon_t \tag{23}$$

Table 1: Model Estimation for the ARIMA Model

Model	Estimated Model	Mle[a]	Se[b]	T-Value	P-Value	Stationary R^2value	R^2 Value
ARIMA(1,1,0)	$(1-0.508B)\nabla z_t = \varepsilon_t$	φ=0.508	0.126	4.028	0.000	0.247	0.824

[a] The maximum likelihood estimates

[b] Standard error

Table 2: Measures of Model Accuracy of the Univariate ARIMA Model

Model	AIC	BIC	MSE	MAE	MAPE	Residual Variance
ARIMA(1,1,0)	-15.02	21.529	45521.864	23212.44	12.178	0.003999

The summary of the fitted univariate ARIMA model of the livebirths series is shown in Table 1. The estimates of the parameters of the proposed univariate models including the associated standard errors, R^2 values, stationary R^2 values, t-values and their corresponding significant values are also reported in Table 1. The associated p-value of the AR parameter shows that its coefficient is statistically significant. The R^2 value for the ARIMA model is 0.824, indicating that over 82% of the total variation in the series is accounted for by the model, and this clearly demonstrate the effectiveness of ARIMA(1,1,0) in modeling the livebirths series. The estimate of the residual standard deviation $\hat{\sigma} = 0.0632$ for the ARIMA model implies that the standard deviation of the residuals is approximately 6% of the level of the series. The estimated coefficients are significant, with t -statistic values well in excess of 2. The overall regression fit, as measured by the R^2 value, indicates a very tight fit.

Measures of diagnostic verification to determine the adequacy of the ARIMA model in representing the underlying process in the live births series were performed. The *acf* of the residuals and Ljung-Box test could be used to check for correlations between successive forecast errors. The *acfs* of the residuals of the ARIMA model show no significant correlation in the residuals for all the lags, except at lag 0, as shown in figure 6(a). The big spike at the beginning is the unimportant lag 0 correlation. The Ljung-Box statistics for each lag up to 10 which are all well above 0.05, indicating non-significance, which is a desirable result as shown in figure 6. The result of the Ljung-Box test on the residuals for the estimated univariate model therefore favours accepting the model as effective models for modelling the persistence in the live births series. The value of the Ljung-Box statistic yielded 0.984 which is not significant. This implies that the model does an excellent job in explaining the observed variation in the series, evidence that the model is correctly specified.

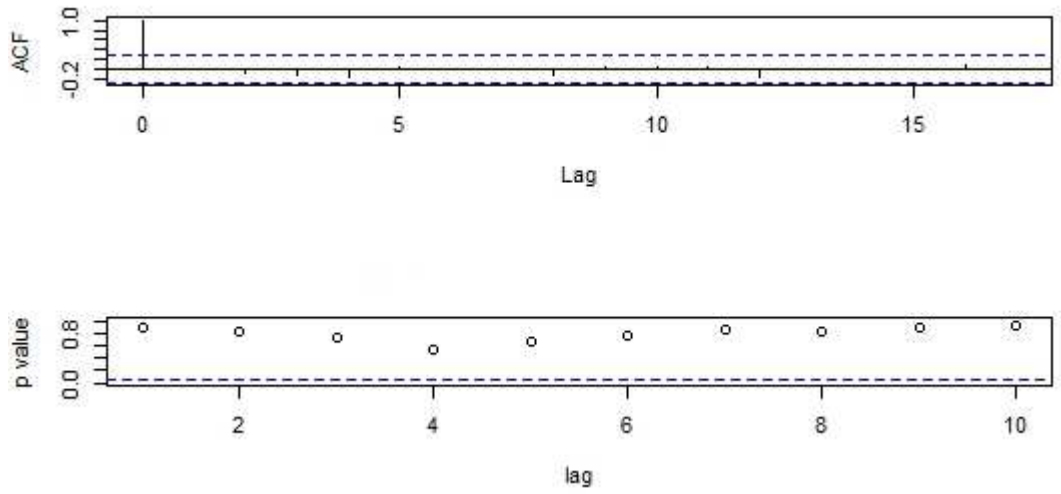


Figure 6: (A) *Acf* of Residuals and (B) P-Values for *Q-Statistic* for ARIMA (1,1,0) Model

Results of the Analysis of the ARIMAX Model

The time plots of the total live births series and the women-of-childbearing-age series are shown in Figures 1a and 1b respectively. There is a noticeable downward trend in the live births series after year 2007. The two series tend to move together in an upward direction over the years, with a natural interaction between them. Starting from the identified univariate ARIMA(1,1,0) model, an ARIMAX model of the live births series (y_t) with one explanatory variable (x_t), number of women-of-childbearing-age is being considered.

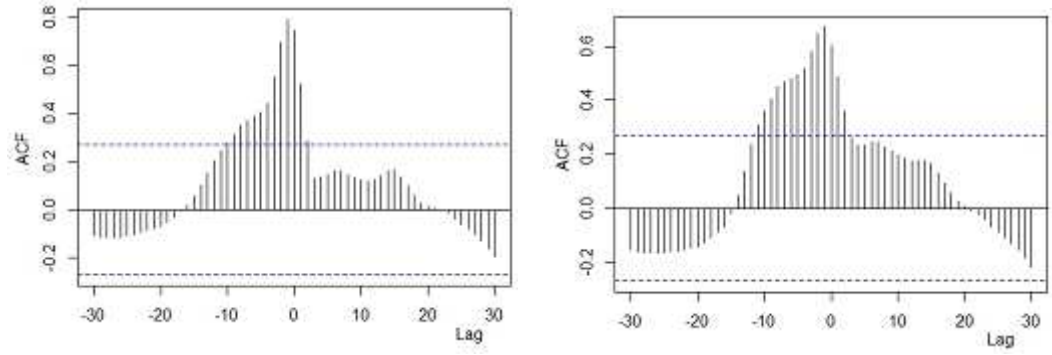


Figure 7: Estimated *Ccf*s for Number of Women-of-Childbearing-Age Vs Livebirths: (A) for Original Series (B) for the Log-Transformed Series

The cross-correlations of the disaggregated series: *women-of-childbearing-age* and *livebirths* and the cross-correlations of the two series after subjecting them to log-transformation are shown in figure 7. It is difficult to read the lags exactly from the *ccf* plots. It is therefore necessary to show the lags and their corresponding correlations. The lags (which is the h in x_{t+h}) and correlation with y_t are shown in Table 3, displaying the correlations from lags -30 to 30. From both figures 7(a) and 7(b) and Table 3, the most dominant cross-correlations occur between lag -1 and lag 0 $(h=-1 \text{ and } h=0)$, and also exhibit strong correlations at positive lags. At lag $h=-1$, $r=0.789$ and at $h=0$, $r=0.744$, which are indications that the strongest correlations occurred at the lag 0 and at lag -1, so that *livebirths* was most strongly correlated with *women-of-childbearing-age* in the current (x_t) and lagged value (x_{t-1}). Inclusion of lagged values of an explanatory variable such as *women-of-childbearing-age* may have a bearing on the

direction of short term trend projections of *livebirths*.

Table 3: Estimated Cross-Correlations between the Series

Lag	Correlations	Lag	Correlations	Lag	Correlations
-30	-0.106	-10	0.281	10	0.126
-29	-0.113	-9	0.314	11	0.121
-28	-0.117	-8	0.350	12	0.123
-27	-0.117	-7	0.372	13	0.142
-26	-0.113	-6	0.386	14	0.161
-25	-0.107	-5	0.405	15	0.165
-24	-0.101	-4	0.443	16	0.136
-23	-0.095	-3	0.552	17	0.099
-22	-0.087	-2	0.694	18	0.059
-21	-0.079	-1	0.789	19	0.029
-20	-0.070	0	0.744	20	0.015
-19	-0.053	1	0.522	21	0.007
-18	-0.030	2	0.286	22	-0.001
-17	-0.006	3	0.128	23	-0.015
-16	0.018	4	0.136	24	-0.038
-15	0.057	5	0.147	25	-0.059
-14	0.104	6	0.161	26	-0.081
-13	0.154	7	0.163	27	-0.103
-12	0.205	8	0.145	28	-0.127
-11	0.245	9	0.134	29	-0.160
				30	-0.196

Starting from the identified univariate ARIMA (1,1,0) model, An ARIMAX model of the explanatory variable lagged by one period (*women-of-childbearing-age*) is formulated. The model captures the effect of the population of women-of-childbearing-age on number of livebirths in Nigeria. The identified ARIMAX model is therefore:

$$(1-\varphi B)\nabla \mathrm{In}(y_t) = \beta_0 \nabla \mathrm{In}(x_t) + \beta_1 \nabla \mathrm{In}(x_{t-1}) + \varepsilon_t \quad (24)$$

For simplicity, let

$$z_t = \mathrm{In}(y_t);\ \ u_t = \mathrm{In}(x_t);\ \ \text{and}\ \ w_t = \mathrm{In}(x_{t-1})$$

Then (24) becomes

$$(1-\varphi B)\nabla z_t = \beta_0 \nabla u_t + \beta_1 \nabla w_t + \varepsilon_t \quad (25)$$

The mathematical expression regress the change in *livebirths* on the change in *women-of-childbearing-age* lagged by one period. This is ARIMAX (1,1,0,1) with one lag of the explanatory variable (x_t). The model could be interpreted as the relationship between the current livebirths and the unpredictable factors in the previous period with lags 0 and -1 for the *women-of-childbearing-age* variable.

Table 4: Estimates of the ARIMAX Model

	Parameter	Estimate	SE	T-Value	P-Value	R^2	Stationary R^2
AR1	φ	0.479	0.127	3.770	0.000	0.939	0.338
∇u_t	β_0	0.118	0.045	2.609	0.012		
∇w_t	β_1	-0.028	0.065	-.0433	0.667		

Table 5: Measures of Accuracy and Measures of Selection of the ARIMAX Model

Mape	Maxape	Maxae	Q	P-Value	Bic Aic
0.864	5.884	0.617	4.374	0.999	-3.383 -29.50

The model contains autoregressive component and coefficients of the explanatory variable at the current and lagged value. The results of the model estimation are reported in Table 4. The estimates of the AR term and that of the lag 0 of the explanatory variable are significant as confirmed by their p-values, while the estimate of the coefficient of the explanatory variable for lag -1 is not significant. The R^2 value of 93.9% shows that a larger portion of the variation in livebirths is explained by the ARIMAX model and thereby suggests a good fit. The positive effect of women-of-childbearing-age (x_t) suggests increase in livebirths. Thus the model postulates that number of livebirths in Nigeria (y_t) depend on number of women-of-childbearing-age (x_t) with one lag. The estimated ARIMAX model is therefore:

$$(1-0.479B)\nabla z_t = 0.118\nabla u_t - 0.028\nabla w_t + \varepsilon_t \quad (26)$$

When the coefficient of the lag of x_t (at lag -1) that is not significant at 95% level was dropped, and the model re-estimated after removing the redundant parameter, the improvements in the root-mean squared error and R^2 are quite negligible.

The best model was determined by comparing values of Akaike Information Criterion and the overall pattern of the models. Comparing the values of the AIC, BIC and MAPE of the univariate ARIMA model in Table 2 and their corresponding values of the measures for ARIMAX model in Table 5, the ARIMAX model was considered to be a more reasonable model for the livebirths series based on the its lower AIC, BIC values and MAPE, compared with the ARIMA model. The criteria of selecting the better model is therefore based on minimization of forecast errors, AIC, BIC and on variance.

The predicted livebirths are plotted with the actual livebirths using the ARIMA and the ARIMAX models in Figure 8. Overall, the predicted livebirths followed a similar pattern of actual cases of livebirths for ARIMAX model better compared with the ARIMA model.

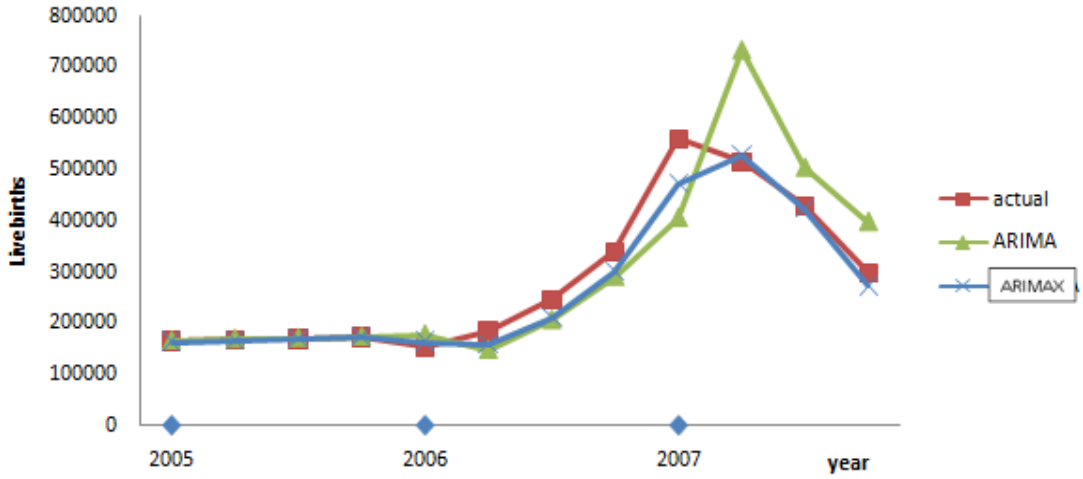


Figure 8: Performance Comparison of the ARIMA and the ARIMAX Models

CONCLUSIONS

ARIMA and ARIMAX time series models are fitted to the live births series, to better understand the data and to predict future trend of live births in Nigeria. The identified univariate time series model is ARIMA(1,1,0), as confirmed by the diagnostic verification of the residuals. The ARIMA model serves as a good representation of the data, captures the time series behaviour of live births, and can be used to provide an adequate basis for forecasting future live births. The only exogenous variable, incorporated into the ARIMAX model is *women-of-childbearing-age* with one lag. The major finding of this study is that, on average the ARIMAX model of live births outperformed the univariate ARIMA model, in terms of the root mean square error. The ARIMA and ARIMAX models fitted the live births series well, and are statistically satisfactory. The models predicted the peak of the disaggregated live births series, and also captured the upward trend of the series.

Forecasting future live births of a population like Nigeria is important, in order to determine forecast demands of this demographic phenomenon, on the various systems in the country such as education, health and economy. The data extracted from vital registration on live births are likely to be inadequate, in a developing nation like Nigeria, due to poor record keeping culture. Despite the possibility of under-reporting of total live births in Nigeria, it is believed that, under-reporting will not substantially alter the proposed time series models, and the forecast of future live births in the country.

ACKNOWLEDGEMENTS

We would like to thank the anonymous reviewers, for their useful comments.

Competing Interest

The authours declare that they have no competing interests.

Authours Contributions

ADB designed, conducted the analysis, interpreted the data and drafted the manuscript. AO made contributions to the acquisition of data, interpretation of data, and also reviewed the article.

REFERENCES

1. *Abraham, B. 1980. Intervention analysis and multiple time series. Biometrika, 67: 73-78.*
2. *Akaike, H. 1974. Markovian representation of stochastic processes and its application to the analysis of autoregressive moving average processes. Ann. Inst. Statist. Math. 26: 363-387.*
3. *Anderson, T.W. 1971. The statistical analysis of Time Series. John Wiley and Sons.*
4. *Bates, D. M., and D. G. Watts. 1988. Nonlinear Regression Analysis and its Applications. New-York: John Wiley and Sons.*
5. *Boot, J. C. G., W. Feibes, and J.H.C. Lisman, 1967. Further methods of derivation of quarterly figures from annual data. Journal of Royal Statistical Society, Series C. Vol. 16(1): 65-75.*
6. *Booth, H. and L.Tickle, 2008. Mortality modelling and forecasting: a review of methods. ADSRI working Paper: no. 3.*
7. *Box, G.E.P. and G.C. Tiao, 1977. A canonical analysis of multiple time series. Biometrika, 64: 355-365.*
8. *Box, G.E.P and L, Haugh, 1977. Identification of dynamic regression models connecting two time series. Journal of the American Statistical Association, 72: 121-130.*

9. *Box, G.E.P., and G.M. Jenkins, 1994. Time series Analysis: Forecast and Control. 3re Edn. Francisco: Holden-Day.*
10. *Brockwell, P. and R. Davis, 1991. Time Series: Theory and Methods. 2nd Edn., New- York: Spinger-Verlag.*
11. *Sneh Saini et al., Application of ARIMA Models in Forecasting Stock Prices, International Journal of Mathematics and Computer Applications Research (IJMCAR), Volume 6, Issue 6, November - December 2016, pp. 1-10*
12. *Chan, W.Y.T and K.F. Wallis, 1978. Multiple time series modelling: another look at the Mink-Muskrat interaction. Applied Statistics, 27:168-175.*
13. *Chatfield, C. 2001. Time series forecasting. London: Chapman and Hall, Inc.*
14. *Chatfield, C. 1989. The analysis of time series, an introduction, 4th Edn., London: Chapman and Hall.*
15. *Chatfield, C. 2004. The analysis of time series, an introduction, 6th Edn., New- York: Chapman & Hall/CRC.*
16. *Chik, Z. 2002. Performance of order selection criteria for short time series, Pakistan Journal of Applied Sciences, 2(7): 783-788.*
17. *Cryer, JD.1985. Time series analysis. Boston:Duxbury Press.*
18. *Dickey, D.A., and W.A. Fuller, 1979. Distribution of the estimators for autoregressive time series with a unit root. Journal of the American Statistical Association, 74: 427-431.*
19. *Diggle, P. J. 1990. Time series, a biostatistical introduction. Oxford: Clarendon Press.*
20. *Dunsmuir, W., and E.J. Hannan, 1976. Vector linear time series models. Advances in Applied Probability, 8: 339-364.*
21. *Fuller WA. 1976. Introduction to statistical time series. New-York: John Wiley.*
22. *Granger, C.W.J. and P. Newbold. 1977. Forecasting Economic Time Series. New-York: Academic Press.*
23. *Hallin, M. 1978. Mixed Autoregressive-Moving-Average multivariate processes with time dependent coefficients. Journal of Multivariate Analysis, 8: 567-572.*
24. *Hannan, E.J. 1970. Multiple Time Series. New-York: John Wiley.*
25. *Hillmer, S.C. and G.C. Tiao, 1979. Likelihood function of stationary multiple autoregressive moving average models. Journal of the American Statistical Association, 74:652-660.*
26. *Hurvich, C.M, and C.L. Tsai, 1989. Regression and Time Series model selection in small samples, Biometrika, 76(2): 297-307.*
27. *Kendall, M. and J.K. Ord. 1993. Time series. 3rd Edn. New-York: Edward Arnold.*
28. *Keyfitz, N. 1966. A life table that agrees with the data. Journal of the American Statistical Association, 63(324): 1253-68.*
29. *Kpedekpo, G.M.K, 1976. Age patterns of fertility in selected African countries. Jimlar Mutane, 1(1): 9-26.*
30. *Ljung, L. 1995. System identification toolbox for use with MATLAB®, User's guide. The Math Works Inc.*
31. *McDonald, J. 1979. A time series approach to forecast Australian total livebirths. Demography, 16:575-602.*
32. *1981. Modelling demographic relationships: an analysis of forecast functions for Australian births. Journal of the American Statistical Association, 76: 782-792.*
33. *McNown, R., and A.Rogers, 1989. Forecasting mortality: a parameterized time series approach. Demography, 26(4): 645-660.*

34. *Moré, J. J. 1977. The Levenberg-Marquardt algorithm: implementation and theory in numerical analysis. In: Lecture Notes in Mathematics, G. A. Watson, Edn. Springer-Verlag.*
35. *National Population Commission [Nigeria], 2008. Report on livebirths, deaths and stillbirths registration in Nigeria, (1994-2007). National Population Commission and ORC/Macro.*
36. *Research in brief. Early childbearing in Nigeria: a continuing challenge. Available from: www.guttmacher.org/pubs/ribs/2004/12/10/rib2-04.pdf. [Accessed December, 2016].*
37. *Saboia, J.L.M. 1977. Autoreggressive integrated moving average (ARIMA) models for birth forecasting. Journal of the American Statistical Association, 72: 264-270.*
38. *Schwartz, G. 1978. Estimating the dimensions of a model. Ann. Statist., 6: 461-464.*
39. *Shryock, HS., J. S.Siegel, and Associates. 1976. The methods and materials of demography, Condensed Edn. London: Academic Press Inc LTD.*
40. *Thompson, PA., W.R. Bell, J.F. Long, and R.B. Miller, 1989. Multivariate time series projections of parametrized age-specific fertility rates. Journal of the American Statistical Association, 84(407): 689-699.*
41. *Tiao, G.C and G.E.P. Box, 1981. Modelling multiple time series with applications. Journal of the American Statistical Association, 76 (376): 802-16.*
42. *U.S. National Centre for Health Statistics, 1967. Physician Handbook on Medical certification: death, fetal death, birth. Public Health Service Publication, Series B 593.*
43. *UNICEF. Available from: www.unicef.org/infobycountry/nigeria_statistics.html, [Accessed September, 2016].*
44. *Upton, G. and I. Cook, 2008. Oxford dictionary of Statistics. New-York: Oxford University Press Inc.*
45. *USAID country health statistical report; Nigeria, 2011. Available from http://www.usaid.org/countryhealthstatisticalreport.html. [Accessed January, 2017].*